# Evaluation of Official Western U.S. Seasonal Water Supply Outlooks, 1922–2002

THOMAS PAGANO AND DAVID GAREN

*National Water and Climate Center, Natural Resources Conservation Service, Portland, Oregon*

SOROOSH SOROOSHIAN

*Department of Civil and Environmental Engineering, University of California, Irvine, Irvine, California*

(Manuscript received 5 September 2003, in final form 30 March 2004)

ABSTRACT

An analysis was conducted of almost 5000 operational seasonal streamflow forecast errors across the western United States. These forecasts are for 29 unregulated rivers with diversity in geography and climate. Deterministic evaluations revealed strong correspondence between observations and forecasts issued 1 April. Forecasts issued earlier in the season were more uncertain yet remained skillful. The average change in forecast performance between January and April was primarily linked to the climatological seasonal cycle of precipitation: regions with climatologically wet winters and dry springs (e.g., California) showed much more forecast improvement between January and April than did regions with dry winters and wet springs (e.g., western Great Plains, Colorado Front Range). Other climatological factors played a secondary role; for example, mixed rain–snow basins in the Pacific Northwest did not show as significant an improvement in skill versus lead time as might otherwise be expected. Mixed trends in 1 April forecast skill were noted since the 1980s, with increased skill in California and Nevada, and a decline in skill in the Colorado River basin. Increased variability in streamflow was also noted across most of the western United States, although this did not appear to be the only factor responsible for trends in forecast skill.

## 1. Introduction

Effective management of limited water supplies is a critical component of the sustainability of populations throughout the western United States. Accurate forecasts of seasonal streamflow volumes assist a broad array of natural resource decision makers. Forecasts with lead times of several months are made possible by the seasonal accumulation and melt of snowpack at moderate and high elevations.

Such water supply outlooks (WSOs) are currently issued jointly by the Natural Resources Conservation Service (NRCS), the National Weather Service (NWS), and local cooperating agencies (such as the Salt River Project in Arizona). They predict the volume of streamflow to pass by a designated point on a stream over a specific period of time, for example, April–September.

NRCS forecasters rely on a statistical principal components regression technique to predict future streamflow using information about current snow water equivalent, fall and spring precipitation, base flow, and climate indices (Garen 1992). In addition, the NWS is increasingly engaged in dynamic simulation of streamflow by initializing a conceptual hydrological model with current soil moisture and snowpack conditions and forcing it with an ensemble of historical meteorological series [ensemble streamflow prediction (ESP); Day 1985]. Human expertise also plays a role in the forecasts from each agency.

Although forecast users are routinely interested in the accuracy of the forecasts (Hartmann et al. 1999), a systematic investigation of the historical performance of the forecasts across the western United States has never been published in a peer-reviewed journal. This paper, then, is an effort to fill this need by evaluating the official WSOs for a select number of basins. It begins with a discussion of the forecasts and data associated with the evaluation. It then continues with a discussion of forecast evaluation methodology, including a review of previous seasonal streamflow forecast evaluation efforts. Next, new evaluations are presented, with a focus on identifying the sources of forecast error as well as the most promising avenues toward forecast improvement. Finally, conclusions are provided based on the analysis herein.

## 2. Selection of basins

The U.S. Geological Survey (USGS) has collected streamflow data at almost 10 000 sites across the western

*Corresponding author address:* Thomas Pagano, National Water and Climate Center, National Resources Conservation Service, 101 SW Main St., Suite 1600, Portland, OR 97204.
E-mail: tpagano@wcc.nrcs.usda.gov

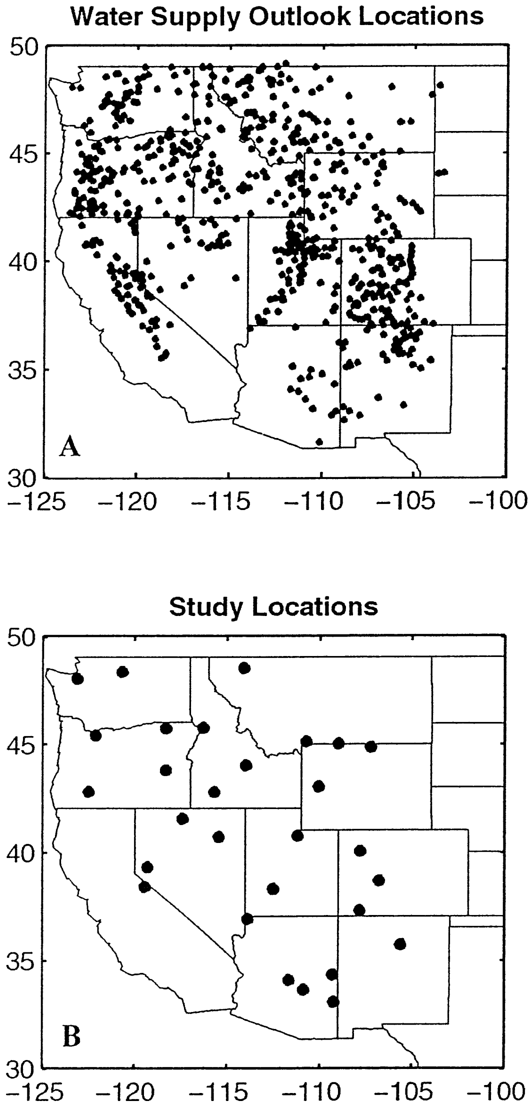


FIG. 1. Maps of (a) the western U.S. seasonal water supply forecast locations as of 2002 and (b) the sites selected for this study.

United States (west of 104° W longitude). Of these sites, WSOs are currently issued for approximately 700 locations (Fig. 1a). These forecast points are generally locations critical to water management operations, for example, inflows to reservoirs or flows at key index sites.

Many forecast points are regulated, in that the observed streamflow is significantly altered by human activity, such as irrigation diversions and reservoir releases. Naturalization of streamflow values to remove human influences is a difficult task, and even the best efforts cannot completely remove all human effects. In reality, there are differences between true natural flow and unregulated flow data (which account for a limited number of measured reservoirs and losses). As a result of these complications, regulated streamflow locations were avoided in this study (see also Dracup et al. 1985). Forecasts in heavily regulated basins are expected to be less skillful than forecasts in unregulated basins. The difference in skill is likely to be inversely proportional to the quality of the naturalized flow data and the extent to which most of the regulations are accounted for.

Slack and Landwehr (1992) identified a subset of Hydro-Climatic Data Network (HCDN) stream gauges as being relatively free of significant human influences and, therefore, appropriate for climate studies. In the continental western United States, there are 481 such points west of 104° W longitude. Excluding Alaska, 151 of the HCDN gauges are currently water supply forecast locations. All of the stream gauges chosen for this study are HCDN locations.

The number of WSO forecast points has increased dramatically over the years. In 1922, a limited set of NRCS forecasts was available in California and Nevada. After the mid-1930s, forecasts increased steadily, adding a net of 11 forecast locations per year, on average. To assess possible trends in forecast accuracy, a long history of forecasts is necessary, limiting the number of basins eligible for analysis. Generally, the selection of basins for this study favored those with a continuous record of forecasts and observations from 1955 to 2002, with stream gauges that are still active today. One may assume that basins with a long period of record of forecasts also have many years of historical streamflow and snowpack data. Data-rich basins have better forecasts than, for example, basins with less than 10 yr of historical streamflow data, where it is difficult to estimate the relationship between snowpack and future streamflow reliably. Although they are very rare, forecasts on ungauged basins are the least reliable.

To ensure relatively complete geographic coverage and a range of basin sizes and types, some basins with shorter forecast records needed to be selected (e.g., the Sandy River near Marmot, Oregon, 1971–2002). Alaskan forecasts were omitted because of their short period of record and compressed forecasting season (i.e., they are only issued in March, April, and May). Finally, the number of basins chosen was limited by the resources available to digitize the historical forecasts manually. Table 1 details the characteristics of the 29 forecast points used in this study, and their locations are shown in Fig. 1b.

## 3. Data

### *a. Observed data*

Monthly streamflow data for the 29 basins were obtained from the USGS online database (available at http://waterdata.usgs.gov/nwis/sw). These monthly values were aggregated into seasonal volumes corresponding to the forecast target period for each basin. Missing observed data were estimated in this study by linear regression between the streamflow for the location of interest and data from nearby stream gauges. Only 1.1%

TABLE 1. Study basins and their characteristics. Latitude and longitude (decimal degrees) are the location of the USGS stream gauge. Sites with italicized names are log transformed. Sites with bold target seasons have shrinking forecast target seasons.

| USGS site name | USGS code | Lat °N | Lon °W | Basin area (km²) | Forecast target season |
|---|---|---|---|---|---|
| Yellowstone River at Corwin Springs, MT | 06191500 | 45.1 | 110.8 | 6794 | Apr–Sep |
| Clarks fork of Yellowstone River, MT | 06207500 | 45.0 | 109.1 | 2989 | Apr–Sep |
| Tongue River near Dayton, WY | 06298000 | 44.9 | 107.3 | 528 | Apr–Sep |
| Pecos River near Pecos, NM | 08378500 | 35.7 | 105.7 | 490 | Mar–Jul |
| East River at Almont, CO | 09112500 | 38.7 | 106.9 | 749 | Apr–Sep |
| Green River at Warren Bridge, WY | 09188500 | 43.0 | 110.1 | 1212 | Apr–Sep |
| White River near Meeker, CO | 09304500 | 40.0 | 107.9 | 1955 | Apr–Sep |
| Animas River at Durango, CO | 09361500 | 37.3 | 107.9 | 1792 | Apr–Sep |
| *Lower Colorado above Lyman Lake, AZ* | 09384000 | 34.3 | 109.4 | 1829 | **Jan–Jun** |
| *Virgin River at Littlefield, AZ* | 09415000 | 36.9 | 113.9 | 13 183 | Apr–Jun |
| *San Francisco River at Clifton, AZ* | 09444500 | 33.1 | 109.3 | 7164 | **Jan–May** |
| *Salt River near Roosevelt, AZ* | 09498500 | 33.6 | 110.9 | 11 153 | **Jan–May** |
| *Verde River below Tangle Creek, AZ* | 09508500 | 34.1 | 111.7 | 15 175 | **Jan–May** |
| Weber River near Oakley, UT | 10128500 | 40.7 | 111.3 | 420 | Apr–Sep |
| Beaver River near Beaver, UT | 10234500 | 38.3 | 112.6 | 236 | Apr–Jul |
| West Walker River near Coleville, CA | 10296000 | 38.4 | 119.5 | 469 | Apr–Jul |
| Carson River near Fort Churchill, NV | 10312000 | 39.3 | 119.3 | 3372 | Apr–Jul |
| Lamoille Creek near Lamoille, NV | 10316500 | 40.7 | 115.5 | 65 | Apr–Jul |
| *Martin Creek near Pardise Valley, NV* | 10329500 | 41.5 | 117.4 | 454 | Apr–Jul |
| *Dungeness River near Sequim, WA* | 12048000 | 48.0 | 123.1 | 404 | Apr–Sep |
| North fork of Flathead River near Columbia Falls, MT | 12355500 | 48.5 | 114.1 | 4009 | Apr–Sep |
| Stehekin River at Stehekin, WA | 12451000 | 48.3 | 120.7 | 831 | Apr–Sep |
| Big Lost River at Howell Ranch, ID | 13120500 | 44.0 | 114.0 | 1166 | Apr–Sep |
| Bruneau River near Hot Spring, ID | 13168500 | 42.8 | 115.7 | 6812 | Mar–Sep |
| Malheur River near Drewsey, OR | 13214000 | 43.8 | 118.3 | 2357 | Apr–Sep |
| Salmon River at Whitebird, ID | 13317000 | 45.8 | 116.3 | 35 094 | Apr–Sep |
| Umatilla River near Gibbon, OR | 14020000 | 45.7 | 118.3 | 339 | Apr–Sep |
| Sandy River near Marmot, OR | 14137000 | 45.4 | 122.1 | 681 | Apr–Sep |
| Rogue River above Prospect, OR | 14328000 | 42.8 | 122.5 | 808 | Apr–Sep |

of the observed streamflow data values needed to be estimated, and, based on the strength of the correlation coefficients of the regression equations, the estimated values are likely to deviate less than 5% from the true values. Therefore, the estimation procedure should not have significantly affected the forecast evaluations. This study assumed that the differences between the forecasts and observations were entirely due to forecast error and were in no part due to the quality of the observations. The USGS considers that approximately 95% of their daily discharge measurements are within 10% of the true value.

Adjusted (unregulated) streamflow data were used only at one location, the Tongue River near Dayton, Wyoming. Although an HCDN location, significant diversions for irrigation occur during the summer months. The accounted-for diversion (Highline Ditch near Dayton, Wyoming, USGS station number 06297500) amounts to 3%–5% of the seasonal streamflow volume. In the 28 other study basins, the NRCS currently calibrates its statistical forecast equations using observed flow data, without any adjustments.

In section 5, the forecast skill improvement is related to the climatological seasonal cycle of precipitation for each of the basins. To support this analysis, spatially distributed climatological average precipitation data were obtained from the Parameter-Elevation Regressions on Independent Slopes Model (PRISM) system. This hybrid statistical–geographical approach blends information about topography and point estimates of precipitation from 1961 to 1990 to derive gridded fields of average annual precipitation at a 4-km resolution for the entire United States (Daly et al. 1994). The project was funded and coordinated by the NRCS National Water and Climate Center and is available on the Internet (online at http://www.ftw.nrcs.usda.gov/prism/prism.html).

### b. Forecast data

The 4841 historical forecasts used in this study were drawn from a variety of existing sources. The primary source of forecasts was paper versions of the historical state ‘‘Basin Outlook Reports’’ and the ‘‘Water Supply Outlook for the Western United States,’’ housed at the NRCS National Water and Climate Center, from which the values were manually digitized. Forecasts after 1990 were available in electronic versions of the same reports. A secondary source of forecasts was the NRCS Forecast Error Analysis Routine (FEAR) electronic database, as used by Shafer and Huddleston (1984). Third, the University of Arizona's Department of Hydrology maintains an electronic archive of water supply forecasts for the Colorado River basin, as they appeared in NWS publications. Finally, paper archives of the publication ‘‘Runoff Forecasts,’’ in the *Western Construction News*

(1947–1954) were used to obtain a very limited number of early forecasts.

The calibration errors of the NRCS water supply forecast regression equations have been used to compute the confidence intervals corresponding to 10%, 30%, 70%, and 90% exceedance probabilities associated with the median (50% exceedance probability) forecasts. These five probability bounds have appeared in NRCS publications since 1989. Before 1986, only median forecasts were published. The publications also generally included values for the historical (e.g., 30 yr) average streamflow for each basin along with the median forecast streamflow as a percent of the historical average. For the remainder of this text, ''forecast'' will refer exclusively to the median forecast. In some publications, this value is referred to as the ''most probable'' forecast, although this is not a statistically rigorous term and is not the preferred terminology.

While the deterministic forecast is the focus of this study, Blanchard (1955) and many others since have demonstrated that an optimal decision maker, for example, an irrigator or reservoir operator, gets more value from a probabilistic forecast than a deterministic forecast. The authors recognize this issue and believe that the deterministic evaluation here is a positive first step toward a fuller probabilistic evaluation. Given that the forecasts were developed using statistical tools, and the forecast distribution width was proportional to expected forecast skill, a probabilistic evaluation should not paint a radically different picture from this analysis. In comparison, simulation models have a well-known tendency to produce overconfident forecasts with narrow forecast distributions in part because they ignore model calibration and data errors (Barnston et al. 2003). Such overconfidence would be penalized in a probabilistic evaluation.

Many forecast points had multiple target seasons. For example, forecasters predicted the April–June, April–July, and April–September flow volume for the Big Lost River in Idaho to serve the needs of different users. Before the 1950s, forecasts almost exclusively had a target period of April–September—the period that corresponds to the irrigation and snowmelt season around most of the western United States. In recent years, to isolate the effects of the relatively unpredictable summer monsoon, Colorado basin forecasts have been for April–July. Other locations may begin snowmelt earlier, such as the Pecos River in New Mexico, which had a forecast target of March–July. Arizona forecasts were unique in that the target period shrank throughout the season. In January, the forecast target was January–May, in February it was February–May, and so on until April–May.

For the purposes of this study, some forecasts' target seasons were changed by multiplying a forecast for a different target season by the ratio of the long-term average flow for the target seasons, as published at the time. For example, multiplying the April–September forecast by the April–July long-term average and dividing by the April–September long-term average creates an estimated April–July forecast. This technique was chosen because it preserved the forecast as a percent of the average, and forecasters commonly developed a forecast for one target period and applied the percent of the average to the other periods. In some situations in this study, concurrent averages for different periods were not available, preventing such a transformation. In these cases, the forecasts were estimated using regressions between the observed streamflow values for the various target periods, excluding the observed flow for the year the forecast is being estimated. In a very limited number of cases, a stream gauge had been permanently moved to a nearby location within the basin, and these forecasts were adjusted to remove the effect of changing the gauge. Of the 4841 unique forecast values, 13% were estimated by one of the means just described. Almost half of the estimated values were on the Weber, Pecos, and Beaver Rivers because of the changing target periods throughout their history. Of these target period changes, that of the Pecos was the most uncertain, with an $R^2 = 0.952$ relationship between March–July and April–September flows. When an estimated value appeared obviously out of line with what a forecaster reasonably would have issued, the forecast was listed as missing.

In many instances, forecasts were cross-checked for consistency among multiple sources. The most common discrepancies were due to keying errors. Discrepancies were resolved on a case-by-case basis, almost always favoring the value that appears in a paper publication. Based on the frequency of discrepancies discovered (and corrected), the authors estimate that at least 99.6% of the forecast values used in this study were identical to the actual forecast. In the instances where forecasts from the NRCS and NWS disagree, the NRCS forecasts were used. Visual inspection of forecast and observation time series and maps ensured that any remaining data entry errors were not gross enough to affect the following analysis significantly.

## 4. Forecast evaluation in the context of previous studies

### *a. History of previous forecast evaluation studies*

Although the water supply forecasting community has long recognized the importance of forecast evaluation, it has also long struggled to find appropriate forecast evaluation measures. The challenge lies in normalizing the forecast errors in some fashion so as to allow for fair comparison between large rivers and minor creeks. Additionally, one must find measures that are understandable and relevant to forecast users.

The early history of the NRCS water supply forecasting program in the 1930s–50s contains many evaluations of individual forecast locations and years (e.g., Paget 1940) similar to those put together for the informal water year–end summary meetings for users that still occur today. Forecast bulletins as early as 1931 contain tables of the previous year's forecasts and observations, and include a text discussion of the performance of the forecasts. Very early evaluations aimed to establish the credibility of the water supply forecasting enterprise.

In the earliest such evaluation, Church (1935) computed the absolute difference between forecast and observation as a percent of long-term average runoff for six basins in Nevada and California. The exceptional result that two-thirds of the forecasts had an error of less than 10% should be tempered by the fact that these forecasts were issued on 15 May, the midpoint of the spring melt period. Additionally, the standard deviation of the observations in this region is typically one-third of the average, indicating that a ''no skill'' forecast every year equal to the long-term average might produce errors of less than 10% one-fourth of the time.

In 1944, the NWS and the NRCS began publishing forecasts independently for many of the same locations. Pressure was put on the agencies to coordinate their forecasting programs to prevent the duplication of effort and to head off the problems that natural resource managers would face when confronted with conflicting forecasts (e.g., *Medford Mail Tribune,* 8 February 1959). The agencies could not agree on the best forecasting method, with the NRCS favoring the use of snow survey data, and the NWS favoring low-elevation accumulated precipitation data. While efforts to institutionalize coordination failed in 1956, regional pockets of coordination continued informally. Most forecast evaluations between 1945 and 1960 were motivated by a desire to show the superiority of the forecasts of one agency over another.

Work and Beaumont (1958) performed a westwide evaluation of NRCS and NWS forecasts, including three tables of analysis arranged by state, basin, and year. Forecast error was defined as the forecast flow divided by the actual flow, expressed as an absolute difference from 100% (after Work 1940). The authors also averaged together the historical forecasts by the various agencies to determine which agency would have benefited by coordination. Finally, the authors presented time series of which agency had the majority of ''best'' forecasts by year and a map of which agency performed best overall at each location. In aggregate, NRCS forecasts were ''better'' than NWS forecasts 11 out of 13 yr of the evaluation. The NRCS forecasts were better at 55% of the locations, although there was no obvious spatial pattern to the performance of the agencies. The authors concluded that snow survey data produce superior streamflow forecasts, because snow is the primary source of water in the western United States.

Kohler (1959), a chief research hydrologist of the NWS, rebutted this study using a different graphical evaluation technique that was in use by Soviet hydrologists at the time. First, the absolute difference between each forecast and observation was expressed as a percentage of the range of the observations. These differences were ranked and plotted as a probability of nonexceedance. A second curve was displayed based on the differences between each observation and the long-term mean, again as a percentage of the range of the observations. If desired, several curves based on the forecasts from different agencies or lead times could have been overlain and the performance compared (as was done in CBIAC 1961, 1964). The area under each error curve was a measure of forecast performance (with a small area being good).

Although not necessarily intuitive to a user, this technique had many appealing aspects in that one could visualize the distribution of errors as well as compare the skill of the forecasts relative to a baseline (such as climatology or always guessing average). While the formulation was not exactly the same, the area between the line for the performance of the forecasts and the line for the performance of the long-term mean approached the spirit of the Nash–Sutcliffe coefficient of efficiency (NS; Nash and Sutcliffe 1970). Using this new forecast evaluation measure, Kohler concluded that the NWS's early season forecasts were far superior to those issued by the NRCS, yet conceded that the differences were slight later in the season.

A lull in forecast evaluation activities followed until the 1980s. Forecast evaluations at individual locations occurred in the research literature, oftentime to compare the historical forecasts against new techniques being developed. It was not until the work of Shafer and Huddleston (1984) that a westwide look at water supply forecast evaluation was revisited. Shafer and Huddleston analyzed a database of close to 50 000 seasonal streamflow forecast errors, representing the complete history of NRCS forecasts (excluding those from Alaska).

Following Church's definition, as opposed to Beaumont and Work's, Shafer and Huddleston calculated the average forecast error for 345 forecast locations and aggregated the results by state and lead time. As expected, the forecast error decreased as the lead time decreased. They also found an exceptional relationship ($R^2 = 0.966$) between the statewide average forecast error and the mean coefficient of variation (the ratio of the standard deviation of the observed flow to the mean), as had Lettenmaier and Garen (1979) in their analysis of streamflow hindcasts several years earlier. In other words, it was easy to incur a 100% forecast error on, for example, the San Francisco River in Arizona, whose observations varied between 17% of average to over 750% of average. It was more difficult to do so on a river such as the Stehekin River in Washington, where the streamflow ranged only between 60% and 150% of average.

Shafer and Huddleston also employed a unique ''skill coefficient'' score, the sum of the absolute differences between the long-term average and the observation in each year, divided by the sum of the absolute errors between the forecasts and observations. A score of 1.0 indicated no skill, and a score of 2.0 indicated that the forecasts were twice as skillful as a climatology forecast. Like Kohler's work, this score was attractive because it enabled a normalized comparison across states. The analysis revealed that 1 April forecasts for the period of record until 1980 were most skillful in Arizona and Washington; fair in Nevada, Idaho, and Wyoming; poor in Colorado, New Mexico, and Utah; and least skillful in Oregon and Montana.

Shafer and Huddleston qualitatively attempted to detect a westwide long-term change in skill, but none was apparent. Individual sites were becoming more skillful, others less skillful. The authors stated that the observations displayed a trend toward increasing variability, and when one subtracted out this effect (based on the above-mentioned analysis), average forecast error decreased ''virtually'' by 2.2% in 1966–80, compared to 1951–65, but decreased ''actually'' only 0.2%. In other words, had the streamflow variability not increased recently, the forecast error would have decreased by 2% more than it actually did. Instead, the forecasters were challenged with a more variable (hence, tougher to forecast) sequence of flows than what occurred in earlier years, and forecast skill suffered.

Schaake and Peck (1985) used a similar score, called the error variance (1 − NS), in an analysis of forecasts during 1947–84 for the inflow to Lake Powell, on the Colorado River in Utah. The authors decomposed the errors into climate-, data-, and model-based error [Lettenmaier and Garen (1979) explored this issue further]. The motivation of the study was to determine the most lucrative avenue for improving streamflow forecasts. Climate-based errors could be addressed by having accurate seasonal forecasts of precipitation and temperature. Data-based errors are rooted in the density of the data-monitoring network, location of sites, and the quality of the data. Improving forecast tools and techniques could reduce model-based errors. Schaake and Peck concluded that almost 80% of the 1 January forecast error was due to the unknown future climate; by 1 April, future climate still accounted for more than 50% of the forecast error. On 1 April, model and data errors were approximately equal and were steady throughout the season.

While water supply forecast evaluation ceased after the mid-1980s, the climate and weather forecasting communities reached new heights of complexity in forecast evaluation. Long-lead climate forecasts were originally issued categorically, for example, ''above normal,'' and the community has no less than 19 categorical evaluation measures at their disposal, with the Heidke skill score (Heidke 1926) as being the most popular in operational circles. With the transition to probabilistic climate forecasts in the 1980s, probabilistic forecast evaluation scores, such as the Brier score (Brier 1950) and ranked probability score (Epstein 1969), gained popularity. The most sophisticated evaluations involve distribution-oriented approaches (reliability and discrimination diagrams, Wilks 1995) and measures of the value of the forecasts to a theoretical optimal decision maker. Regrettably, the robustness and scientific rigor of forecast evaluation techniques are inversely proportional to their accessibility and understandability by the lay forecast user. The current challenge is in linking the forecast evaluation to the user in a meaningful and relevant manner (Hartmann et al. 2002).

### b. *Current forecast evaluation*

Although past NRCS forecast evaluations focus on the average percent error, and this measure is the most easily understandable by users, this study did not use it. It primarily measures the local variability of the observations and not the value added by the forecaster. Instead, the forecasts were judged by the NS score:

$$\mathrm{NS} = 1 - \frac{\sum_{i=1}^{N} (f_i - o_i)^2}{\sum_{i=1}^{N} (\overline{o} - o_i)^2}, \tag{1}$$

where $f_i$ and $o_i$ are the forecast and observations in year $i$ for a collection of $N$ years, and $(\overline{o})$ is the mean of the observations of $N$ years. An NS of 1 is perfect, 0 indicates no skill over the always guessing average, and values less than 0 mean negative skill. In essence, this score is one minus the mean squared error of the forecasts divided by the variance of the observations. It is important to note that this skill score already accounts for changes in the forecast skill associated with changes in the variability of observations. During periods of high variability, the potential for greater forecast error is offset by increased error in the ''no skill'' baseline forecast of the guessing average.

Although not necessary, it is useful to avoid situations of heteroscedastic error, such as where the forecast error is typically greater during high flows than low flows. Among those locations where the seasonal flow volumes have skewness greater than 1.0, the natural logarithm is applied to the forecast and observed seasonal totals before analysis (see Table 1). Otherwise, individual large floods would dominate the analysis, resulting in an evaluation reflecting the behavior of the streamflow in a few years rather than the quality of the forecaster on the whole. For example, the 1 January 1993 squared forecast error for the San Francisco River in Arizona was almost 110 times the median squared forecast error over the period of record even though it was, by far, the wettest forecast ever issued for this location. Instead, analysis of log-transformed flows provides more information about the performance of forecasts across a range of streamflow conditions. In this study, the transformation increased the skill scores (e.g., San Francisco River January NS was 0.64 if the data were transformed, and 0.54 if not). This transformation shifted the forecast evaluation emphasis to drought, which is the primary concern of NRCS agricultural customers in semiarid regions (as opposed to NWS customers concerned with the protection of lives and property from floods). Further, droughts (FEMA 1995) cause approximately 3–4 times the annual economic damages of floods (Myers 1997), and, therefore, any impact-oriented forecast evaluation may prefer to give more emphasis to performance during dry years. Although this study did not attempt to link forecast accuracy to user benefits, the log transformation of skewed flows increased the user relevance of the evaluation.

## 5. Evaluations

The evaluation of the historical forecasts as a function of location and lead time is presented below. Changes in skill over the period of record were investigated, although quantitative detection or diagnosing of trends in forecast performance were not attempted. A limited number of factors that shaped forecast performance in a general sense were identified and examined, with examples from individual years.

### *a. Forecast lead time*

Figures 2a–c are maps of the NS score for the forecasts issued in the most recent 20 yr, 1983–2002. The top and middle maps indicate the performance of the forecasts issued on 1 January and 1 April, respectively. The size of the circle reflects the skill of the forecasts, large being preferable over small. The outermost circle is a reference to perfect skill, and an empty circle indicates zero skill. The hollow circle over the Sandy River in Oregon in the top panel indicates that 1 January forecasts had a slightly negative skill. The bottom panel reflects the change in forecast performance (NS) between 1 January and 1 April. Large inner circles indicate great forecast improvement, and the hollow circle in Arizona shows a decline in skill for the Verde River in Arizona, 1 April forecasts compared to those issued on 1 January. The outer circle is a reference for a change in NS equal to 1.0.

During 1983–2002, the most skillful 1 April forecasts were issued for the Salt River in Arizona, West Walker River in California, and Little Colorado River in Arizona, whereas the least skillful forecasts were for the Umatilla River in Oregon, White River in Colorado, and Sandy River in Oregon. The most improvement in skill between January and April occurred for the West Walker River in California, Carson River in Nevada, and Martin

→

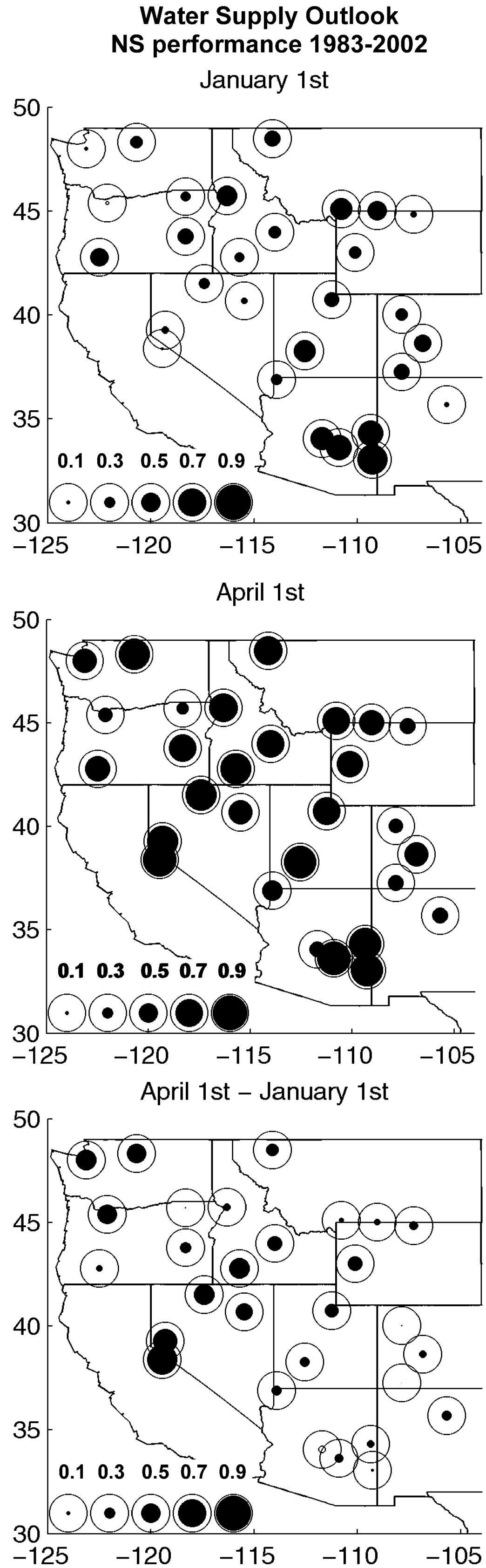


FIG. 2. Skill of 1983–2002 water supply forecasts issued (top) 1 Jan and (middle) 1 Apr. (bottom) Forecast improvement between Jan and Apr is shown. Large filled circles indicate high skill or great improvement.

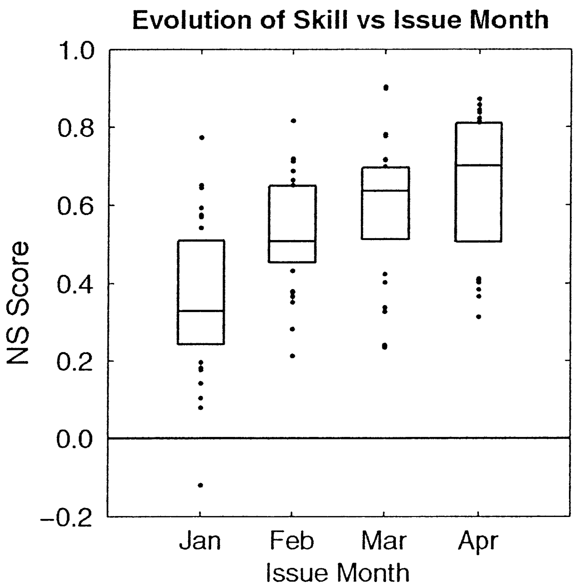


FIG. 3. Box diagram of forecast skill versus issue month for the 29 basins during 1983–2002. The box has lines at the lower-quartile, median, and upper-quartile values. Individual sites outside this range are shown as dots.

Creek in Nevada, and the least improvement occurred for the Verde River in Arizona, Animas River in Colorado, and White River in Colorado. The westwide average NSs in January through April from 1983 to 2002 were +0.36, +0.53, +0.59, and +0.65, respectively.

Figure 3 shows the westwide average forecast skill versus issue month. Skill was lowest but generally positive in January and steadily improved throughout the season. This result is intuitive in that in January the character of the seasonal precipitation has yet to reveal itself. For many locations in the western United States, snowpack is at its peak on or around 1 April, and there are fewer opportunities for dramatic changes in the amount of available water in the basin.

This is not to say that significant changes cannot occur after 1 April. A notable example of this occurred in the Colorado River basin in 1983 (Rhodes and Dracup 1984). Until April 1983, snowpack was near average, and the median forecasted inflow to Lake Powell was similarly near average (109%). An exceptionally cold and wet spring ensued, followed by a rapid warming. The observed April–July flow, at over 210% of average, overwhelmed the already full reservoir system. Perhaps a less well known example, the Animas basin, had extremely low snowpack on 1 April 1999 after an exceptionally warm March, and the median streamflow forecast was approximately 40% of average. Near-record rain fell in April–May, and the snowmelt pulse volume was near to above average. A monsoon of unprecedented strength, however, produced summer floods. April–September streamflow totals were close to 140% of average. Although not one of this study's basins, Ponil Creek near Cimarron, New Mexico (USGS station number 07207500), to the southeast, had a forecast of 20% of average in the same year, and the eventual March–June flow was over 370% of average. The five largest 1 April errors in the database, defined as the squared difference between the forecast and observed, divided by the long-term observed variance, were (beginning with the largest first) as follows: Animas River in Colorado, 1999; Verde River in Arizona, 1988; Sandy River in Oregon, 1981; Bruneau River in Idaho, 1963; and East River in Colorado, 1957. These were all due to low snowpack conditions followed by exceptional storms.

While these examples represented large underforecasts, causes for overforecasts tended to be more complex. The two largest forecast overestimates (Yellowstone River in Montana, 1949; Weber River in Utah, 1936) were due to moderate snowpack being followed by hot, dry, windy weather. The Lamoille River in Nevada, 1943 error, however, was a collision of factors. In this case, forecasters overestimated the effects of a high water table, and this very small high-elevation basin seemed to be in a dry microclimate surrounded by heavy snows. The streamflow data were also questionable because a series of floods had recently occurred, possibly affecting the instruments and rating table (Church and Boardman 1944).

Forecast error in any given year is strongly related to the character of precipitation that falls subsequent to the forecast issue date. For example, a basinwide time series of spring and summer precipitation explained more than 60% of the interannual variance in the 1 April forecast errors at the Weber River in Utah (Pagano 2004). Similarly, one might expect the average increase in forecast skill between January and April to be proportional to the percentage of precipitation that typically falls in January, February, and March. The following analysis determined if the operational forecasts displayed this characteristic.

The stream gauge coordinates and a 1-km digital elevation model (HYDRO1k; see information online at http://edcdaac.usgs.gov/gtopo30/hydro/namerica.asp) were used within a geographic information system to delineate the 29 basins of this study. The long-term climatological average PRISM precipitation for each month was calculated within each basin's boundaries. The January–March total precipitation was then divided by the total seasonal average precipitation, beginning in January. For example, if the water supply forecast target season was April–September, ''seasonal precipitation'' meant January–September (or January–July if the target season was April–July). A high value indicated that a large portion of the streamflow-relevant precipitation typically fell in January–March, and the spring was a climatologically dry period. A moderate value (~40%) indicated a relatively flat seasonal cycle to precipitation, and a very low value indicated that most precipitation tended to fall in the spring and summer. Arizona rivers were withheld from this analysis, because their target

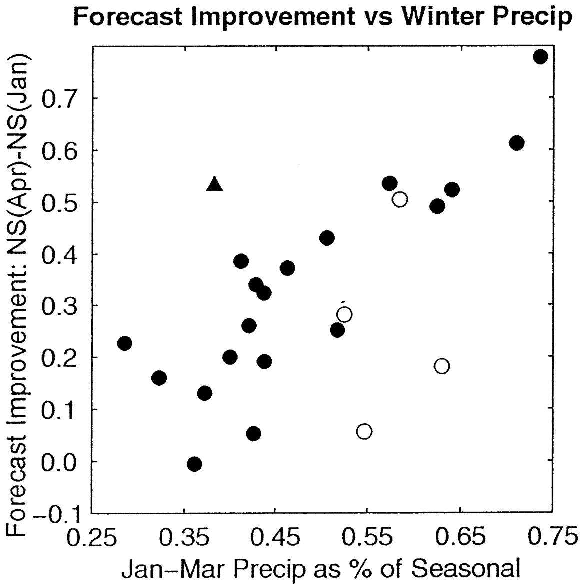


FIG. 4. Forecast skill improvement between Jan and Apr 1983–2002, as a function of the climatological average Jan–Mar precipitation, relative to the full seasonal (Jan through the end of the forecast target period) precipitation. See text for explanation of symbols.

period shrank as the season progressed, and a smaller target was not necessarily easier to hit.

The strength of the correlation between average forecast improvement and the climatological cycle of precipitation in Fig. 4, as expected, was relatively strong ($R^2 = 0.41$). The river with the greatest improvement in skill between January and April was the West Walker River in California, because 46% of the annual precipitation falls in January–March, compared to only 17% that falls in April–July. In comparison, the Tongue River in Wyoming showed little improvement in skill by April, in part because April–June in the Plains states is usually the wettest time of year. The Animas River in Colorado showed the least improvement of any basin outside Arizona due to the aforementioned 1999 event and because the basin is under the influence of the summer monsoon; the recent operational switch of forecast target periods from April–September to April–July may address some of this problem. The outlier (filled triangle) in the upper left corner of the diagram is the Bruneau River in Idaho; as the target season began in March, observed March flow for this location was known in real time. On 1 April, the forecaster was given an artificial advantage because part of the target season was in the past and, therefore, known with complete confidence.

The other exceptions to this rule, in the lower right corner of the diagram as hollow circles, were the four rivers in Oregon that did not improve as much as expected versus lead time. Most snowmelt-dominated basins around the western United States have a strong seasonality in streamflow, with low base flow from September to March, a rise in late spring, a peak in summer, and recession in the fall. For the East River, only 5% of the January–September streamflow typically occurs in January–March. In contrast, Oregon basins experience a mix of rain and snow and display "peaky" hydrograph behavior during the winter. On average, 47% of the January–September flow on the Sandy River typically occurs in January–March (43%, 42%, and 37% for the Umatilla, Malheur, and Rogue Rivers, respectively). It is possible for a large snowpack in February to be wasted away by March rains and run off before the April–September forecast target period begins. For example, on 1 February 1996, the Sandy River watershed had a near-average snowpack and near-average streamflow forecasts. Warm temperatures and heavy rains caused major flooding and resulted in February's streamflow being 260% of average. By 1 April, the snowpack was 50%–60% of average, and the eventual April–September flow was among the driest third of record. The special challenge of seasonal streamflow forecasting in Oregon is evident.

### b. *Trends in forecast skill*

Not only are operational agencies under pressure to use scientifically sound judgement to avoid large forecast errors, they are also under pressure to assess the value of the forecasts and to display improvements in skill associated with the adoption of new technology. Unfortunately, this is not necessarily an easy task. For example, in the context of discussions of changes in forecast skill, Lettenmaier (1984) deflated hopes of measuring the expected 6% increase in forecast accuracy (and the multimillion dollar annual benefit to the economy) associated with satellite snow cover information. Lettenmaier showed that it would take more than half a century to accumulate enough forecasts to detect such a small accuracy trend with confidence. Mindful of the concerns about the decades of forecasts necessary to detect small changes in skill, this study found mixed trends similar to those of Shafer and Huddleston (1984).

For each location, the NS was evaluated for forecasts of various lead times within a 20-yr moving window. Some locations (Figs. 5a–b), such as the Big Lost River in Idaho, displayed monotonic improvements in forecast skill, whereas other locations, such as the Pecos River in New Mexico, experienced recent drops in skill (the Pecos peaked in 1966–85). Most basins had slowly improving forecasts until one or two major forecast anomalies ruined the moving average forecast performance. For example, forecast errors in 1983 and 1999 dominated much of the recent analysis in the upper Colorado River basin.

Figure 6 shows the westwide average performance of the forecasts as a function of lead time. April forecasts improved through the early part of the record, reached a relative maximum in the 1970s, and declined shortly after. Longer lead time forecasts steadily improved throughout the period of record.

One might suspect that the downward trend in late-

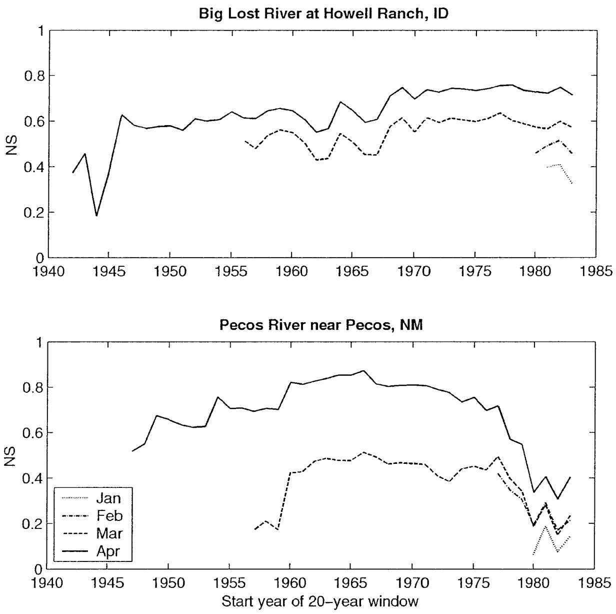


FIG. 5. Evolution of forecast skill over time, as measured within a 20-yr moving window for various forecast issue months (legend shown in lower left).

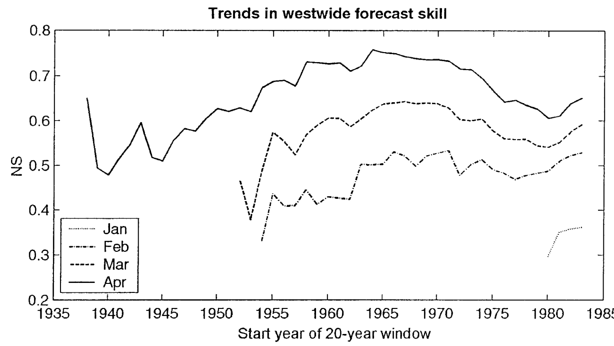


FIG. 6. Westwide 20-yr moving average forecast skill versus year and issue month (see legend lower left). At least 6 of the 29 basins must have valid data for a value to be shown.

season skill was artificial. The average in the early part of the record could have been high because of the inclusion of high-skill basins with a long period of record, such as that of the Walker and the exclusion of low-skill basins such as that of the Sandy River. To evaluate this possibility, each site's 20-yr moving window NS for 1 April was normalized into an anomaly: $NS_i' = NS_i - \overline{NS}$, where $\overline{NS}$ was the average of the moving window NS values over the period of record. This transformation facilitates easy comparison of relative trends in skill across sites. These anomalies are displayed in Fig. 7. The start year of the 20-yr moving window is on the $x$ axis, whereas the site number from Table 1 is on the $y$ axis. At each location and period, an empty circle indicates a positive skill anomaly, meaning that performance at the location during this period was better than other periods in this site's history (a "streak"). A filled circle denotes a negative anomaly or performance worse than other times in the past (a "slump"). The

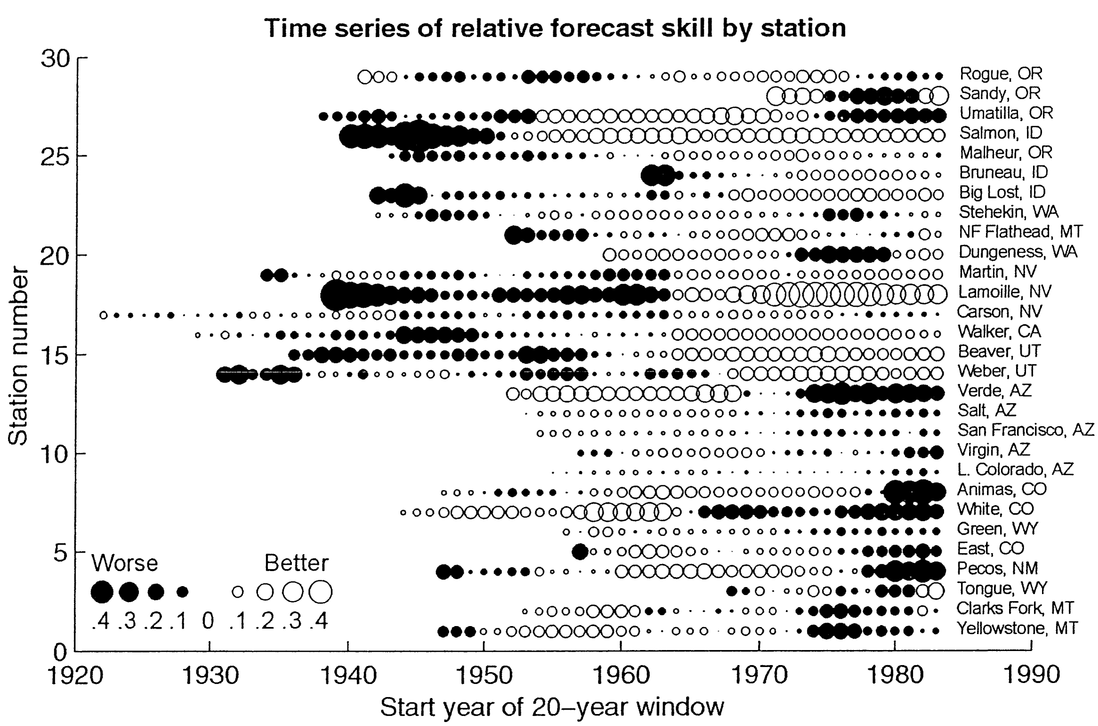


FIG. 7. The 1 Apr forecast skill anomalies by location versus year. Large hollow circles are preferable over large filled circles.

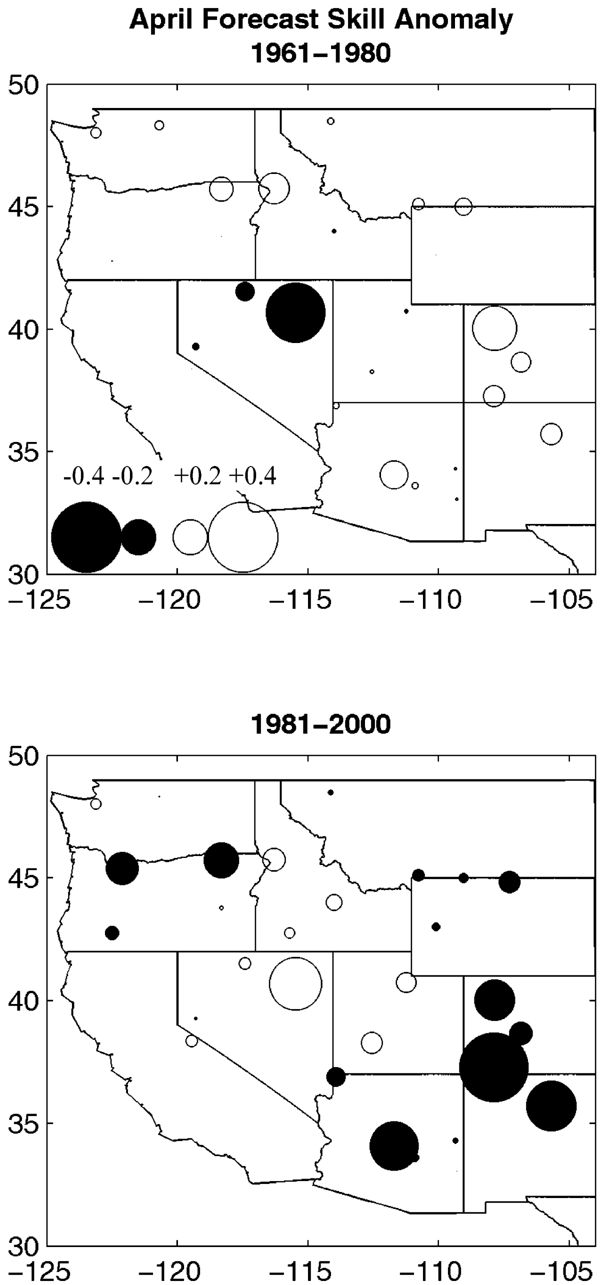


FIG. 8. Maps of (top) 1961–80 and (bottom) 1981–2000 1 Apr forecast skill anomalies. Large hollow circles are preferable over large filled circles.

size of the circle reflects the magnitude of the anomaly, with a larger circle meaning a greater magnitude. Figure 8 is a map of 1 April forecast skill anomalies for 1961–80 (top) and 1981–2000 (bottom). As with Fig. 7, large filled circles indicate a slump in performance.

Forecasts in California, Nevada, and southern Idaho (sites 16–19 and 23–24) performed relatively poorly from the 1940s to the mid-1980s, after which time forecast skill was at its greatest of any time in the past. Many other sites, however, had a relatively high skill during the 1950s–80s and a decline afterward. This decline was most pronounced in the Colorado River basin (sites 5–8) and Arizona (sites 9–13). Across the western United States, the period 1968–87 marked a relative high point in forecast skill, with 26 of the 29 sites used in this study reporting positive NS anomalies (skill higher than other periods). Less than 10 yr later, 1980–99 marked a relative low point in modern forecast skill, with only 10 of 29 sites reporting positive NS anomalies. Skill climbed modestly since, with 13 sites reporting positive scores during 1983–2002. Therefore, the downward trend in 1 April skill in Fig. 6 was not artificial and was not an artifact of mixing forecast time series of different lengths.

Without complete metadata about changes in forecast procedures and data collection, it is difficult to diagnose the trends in forecast skill. The recent decline in skill coincided with many changes in the forecasting environment. The early 1980s saw a major restructuring of forecast facilities within the NRCS, from being state based to being centrally located (Barton 1983). The NRCS forecasting staff in 2003 was only one-third of its size in 1980. Garen (1992) developed a significantly different statistical forecasting technique that found wide use after the early 1990s. Some NWS offices adopted seasonal simulation modeling of streamflow and the ESP system in the late 1970s.

The automation of snow courses was phased in over the 1980s with the advent of the Snow Telemetry (SNOTEL) network, which was an improvement but also a discontinuity in data collection technology. Changes in land use, small water impoundments, and undocumented diversions could have affected the future representativeness of historical flow. It is unknown whether the current snow-based forecasting equations are representative under a climate that is warming and oscillating on decadal time scales. The forecasts were objectively based initially, but the published values were oftentimes adjusted using nonquantifiable and nonreproducible human professional judgment. Both the statistical procedures and the human operators changed over the history of forecasting, as will they change in the future (although it would be misguided to ascribe a significant rise or fall in skill to an individual person, given the many parties involved with creating a forecast). Some, all, or none of these factors may have shaped how forecast skill evolved recently. Preliminary analysis (not shown) pointed to climate variability as the most influential factor on skill trends, while snow data quality had secondary importance. The changes in human forecasters or forecasting technology had small, if at all detectable, effects on skill trends. This issue is explored further in Pagano (2004).

### *c. Trends in streamflow variability*

Researchers have postulated that streamflows may become more variable under climate change, with an in-

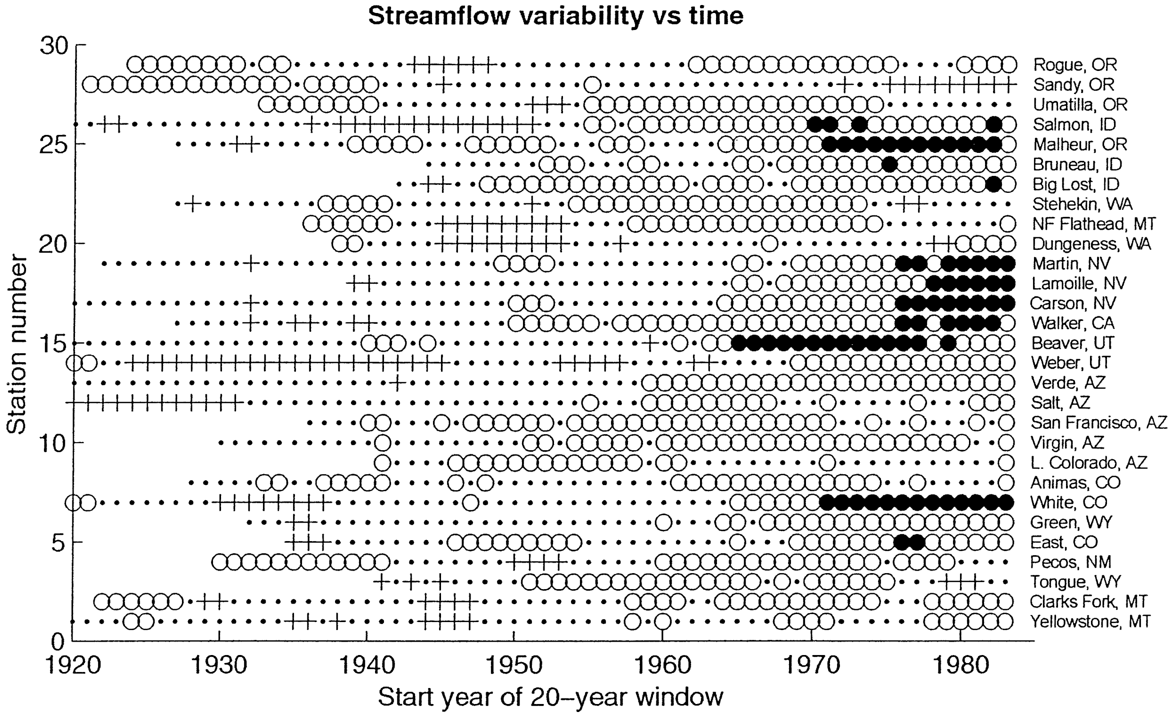


FIG. 9. Time series of moving window–observed streamflow variability. Compare with Fig. 7. Circles indicate more variability than the period of record, dots indicate less variability. Variability declines significant at the $p = 0.1$ level are crosses, and statistically significant variability increases are shown as filled circles.

creasing frequency of high and low flows (Trenberth 1999). The operational community expects forecast skill to decrease as streamflow variability increases; for example, Shafer and Huddleston (1984) indicated that streamflows had become more variable, and this masked improvements in average forecast error. The reader is reminded, however, that the evaluation skill score used in this study is already normalized by the variability of the observations, so this explanation for recent declines in forecast skill does not seem to be sufficient.

Nonetheless, Fig. 9 shows the trends in forecast target period streamflow variability versus time. The ratio of the local 20-yr variance to the period of record variance after 1900 is plotted. A dot indicates a ratio less than one, namely periods of relatively low variability. Crosses are plotted where the ratio passes a variance $F$ test at $p = 0.1$ (i.e., the decline was statistically significant). Increased variability is shown as circles, and statistically significant variability increases are filled circles. The threshold for statistical significance varied by location but was generally in the range of variance less than 60% or greater than 150% of the period of record variance. Figure 10 shows this information in map form for 1961–80 (top) and 1981–2000 (bottom). Shown is the ratio of the variance of flows in the 20-yr period compared to the period of record. Symbols pointing upward indicate where variance increased, and those pointing downward are where variance decreased. If the symbol is filled the change was statistically significant.

During the period 1945–1964, all but one of 29 sites across the western United States had less variability than the period of record, 10 sites significantly so. Over 1971–90, 27 of 29 sites were more variable than their period of record. For the most recent 20 yr, 24 of 29 sites were more variable. This peak in variability was focused in the intermountain western United States, centered on California, Nevada, southern Idaho, and the northern half of the Colorado River basin. When using a skill score that normalizes by variability in the observations, it is difficult to attribute downward trends in forecast skill to recent increases in observed variability. In the Great Basin (the closed basins of Nevada and western Utah) and California, both forecast skill and streamflow variability have risen. In the Colorado River basin, streamflow variability rose, but forecast skill declined. This analysis did not account for changes in springtime precipitation variability or the autocorrelation between winter and spring precipitation, which may have affected 1 April forecast skill. For example, Pagano (2004) found that the recent forecast skill declines resembled, in time and space, the unprecedented rise in interannual variability in spring/summer precipitation in the Southwest and Pacific Northwest since the mid-1980s.

## 6. Conclusions

This study analyzed the operational seasonal water supply forecasts for a diverse set of basins evenly dis-

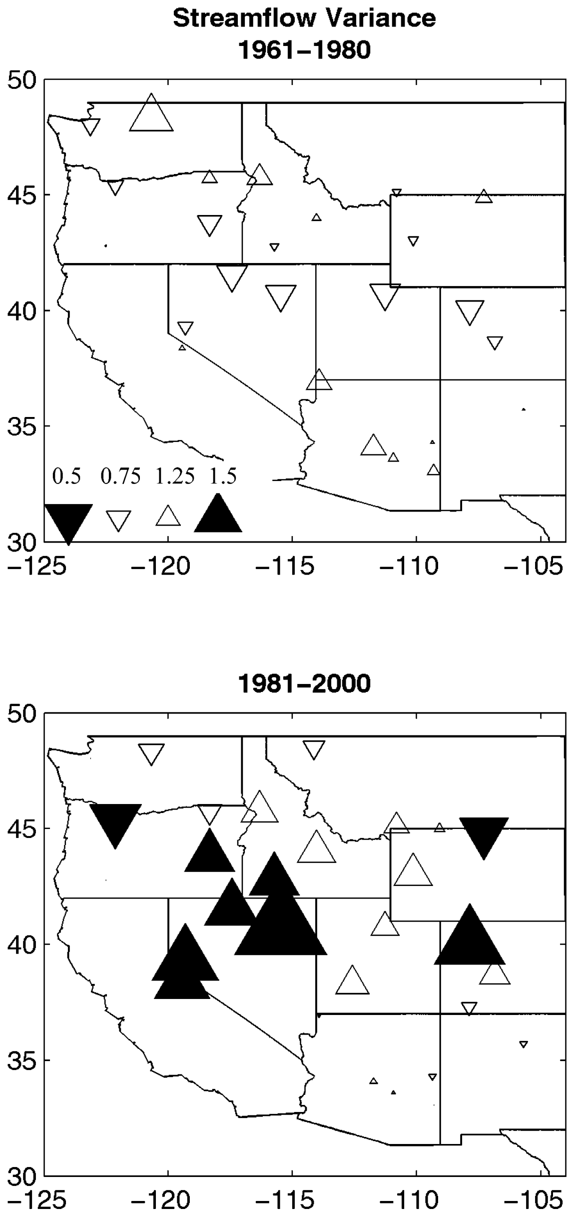


FIG. 10. Observed streamflow variance for (top) 1961–80 and (bottom) 1981–2000 as a ratio of the period of record variance. See text for symbol definitions.

tributed around the western United States. For the most recent 20 yr, all of the basins had skillful 1 April forecasts, and all but one basin had skillful 1 January forecasts. This suggests that forecasters may be able to produce skillful forecasts before 1 January for a subset of basins. This advancement is technically feasible because sufficiently long records of pre-January snow data are now available through the SNOTEL network, and credible seasonal climate forecasts and indices now exist.

This study found that the increase in skill between January and April was directly related to the proportion of seasonal precipitation that typically falls in January–March. Therefore, regions such as California, with relatively compressed precipitation seasons, saw dramatic increases in forecast skill between January and April. The exceptions to this rule were mixed snow–rain basins in the Pacific Northwest. Even if January–March typically accounted for a large percentage of the seasonal precipitation, if winter streamflow was a significant portion of the annual streamflow, seasonal skill improvement was diminished.

Therefore, a measure of expected forecast skill could be derived from climatological parameters and compared with the jackknife calibration error of the forecasting equation to detect possible overfitting. The jackknife technique involves calibration of an equation on all but one historical year of data and then using the equation to predict the single year that was removed. This process is repeated leaving out each historical year in turn until a full set of predictions is obtained. The jackknife approach usually provides a reliable estimate of expected operational forecast skill unless too few years are used in the calibration.

Trends in forecast skill were noticed in both absolute and relative terms. At this time it is difficult to diagnose the cause for the relative maximum in westwide forecast skill from the 1950s to the 1970s, the decline in skill that followed, and the recent trend toward recovery. While streamflow variability rose across the western United States, the spatial pattern of changes in variability and changes in skill did not completely match. Specifically, forecast skill declined in the Colorado River basin and Arizona since the 1980s. Skill increased in California, Nevada, and western Utah, yet, streamflow variability increased over the same period for almost every western U.S. basin between northern Arizona and central Idaho. Future research may reveal the underlying causes for the changes in forecast skill, such as the recent rise in spring precipitation variability or the impacts of regional warming on the relationship between snowpack and subsequent streamflow.

*Acknowledgments.* This study was part of the first author's doctoral research at the Department of Hydrology and Water Resources at the University of Arizona. Partial support for this research came from the CLIMAS project (National Oceanic and Atmospheric Administration Office of Global Programs Grant NA 86GP0061) and the American Geophysical Union Horton Research award.

REFERENCES

Barnston, A. G. B., S. J. Mason, L. Goddard, D. G. DeWitt, and S. E. Zebiak, 2003: Multimodel ensembling in seasonal climate forecasting at IRI. *Bull. Amer. Meteor. Soc.,* **84,** 1783–1796.

Barton, M., 1983: Reorganization of the USDA–Soil Conservation Service—Snow survey and water supply forecast activity. *Proc. Western Snow Conf.*, Vancouver, WA, Western Snow Conference, 151–154.

Blanchard, F. B., 1955: Operational economy through applied hydrology. *Proc. Western Snow Conf.,* Portland, OR, Western Snow Conference, 35–48.

Boardman, H. P., G. G. Devore, and L. Stanford, 1940: Nevada cooperative snow-surveys—Eastern slope, central Sierra Nevada: Comparison of forecast and actual results, 1939, *Trans. Amer. Geophys. Union,* Part I-B, 99.

Brier, G. W., 1950: Verification of forecasts expressed in terms of probability. *Mon. Wea. Rev.,* **78,** 1–3.

CBIAC, 1961: Review of procedures for forecasting inflow to Hungry Horse Reservoir, Montana. Columbia Basin Interagency Committee Rep. 16, 69 pp.

——, 1964: Derivation of procedure for forecasting inflow to Hungry Horse Reservoir, Montana. Columbia Basin Interagency Committee Rep. 20, 63 pp.

Church, J. E., 1935: Principles of snow surveying as applied to forecasting stream flow. *Agric. Res.,* **51,** 97–130.

——, and H. P. Boardman, 1944: Effect of high water-table in exaggerating streamflow: Further analysis of unusual year 1942–43 and analysis for year 1943–44 on Upper Humboldt River, Nevada. *Trans. Amer. Geophys. Union,* **25,** 96–101.

Daly, C., R. P. Neilson, and D. L. Phillips, 1994: A statistical–topographic model for mapping climatological precipitation over mountainous terrain. *J. Appl. Meteor.,* **33,** 140–158.

Day, G. N., 1985: Extended streamflow forecasting using NWSRFS. *J. Water Resour. Plan. Manage.,* **111,** 157–170.

Dracup, J. A., D. L. Haynes, and S. D. Abramson, 1985: Accuracy of hydrologic forecasts. *Proc. Western Snow Conf.,* Boulder, CO, Western Snow Conference, 13–24.

Epstein, E. S., 1969: A scoring system for probabilities of ranked categories. *J. Appl. Meteor.,* **8,** 985–987.

FEMA, 1995: Mitigation directory. *National Mitigation Strategy: Partnerships for Building Safer Communities,* Federal Emergency Management Agency, 40 pp.

Garen, D. C., 1992: Improved techniques in regression-based streamflow volume forecasting. *J. Water Resour. Plan. Manage.,* **118,** 654–670.

Hartmann, H. C., R. Bales, and S. Sorooshian, 1999: Weather, climate, and hydrologic forecasting for the Southwest US. University of Arizona Climate Rep. Series CL2-99, 172 pp.

——, T. C. Pagano, S. Sorooshian, and R. Bales, 2002: Confidence builders: Evaluating seasonal climate forecasts from user perspectives. *Bull. Amer. Meteor. Soc.,* **83,** 683–698.

Heidke, P., 1926: Berechnung des erfolges und der gute der windstarkevorhersagen im Sturmwarnungsdienst (Measures of success and goodness of strong wind forecasts by the gale-warning service). *Geogr. Ann.,* **8,** 301–349.

Kohler, M. A., 1959: Preliminary report on evaluating the utility of water supply forecasting. *Proc. Western Snow Conf.,* Reno, NV, Western Snow Conference, 26–33.

Lettenmaier, D. P., 1984: Some issues in assessing the accuracy of hydrologic forecasts. *Proc. A Critical Assessment of Forecasting in Water Quality Goals in Western Water Resources Management*, Bethesda, MD, American Water Resources Association, 106–116.

——, and D. C. Garen, 1979: Evaluation of streamflow forecasting methods. *Proc. Western Snow Conf.,* Sparks, NV, Western Snow Conference, 48–55.

Myers, M. F., 1997: Trends in floods. *Proc. Workshop on the Societal and Economic Impacts of Weather*, Boulder, CO, National Center for Atmospheric Research, 77–86.

Nash, J. E., and J. V. Sutcliffe, 1970: River flow forecasting through conceptual models. Part I: A discussion of principles. *J. Hydrol.,* **10,** 282–290.

Pagano, T. C., 2004: The role of climate variability in operational water supply forecasting for the western United States. Ph.D. dissertation, University of Arizona, 262 pp.

Rhodes, S., D. Ely, and J. Dracup, 1984: Climate and the Colorado River: Limits to management. *Bull. Amer. Meteor. Soc.,* **65,** 682–691.

Schaake, J. C., and E. L. Peck, 1985: Analysis of water supply forecast accuracy. *Proc. Western Snow Conf.,* Boulder, CO, Western Snow Conference, 44–53.

Shafer, B. A., and J. M. Huddleston, 1984: Analysis of seasonal volume streamflow forecast errors in the western United States. *Proc. A Critical Assessment of Forecasting in Water Quality Goals in Western Water Resources Management*, Bethesda, MD, American Water Resources Association, 117–126.

Slack, J. R., and J. M. Landwehr, 1992: Hydro-Climatic Data Network: A U.S. Geological Survey streamflow data set for the United States for the study of climate variations, 1974–1988. U.S. Geological Survey Rep. 92–129, 193 pp.

Trenberth, K. E., 1999: Conceptual framework for changes of extremes of the hydrological cycle with climate change. *Climate Change,* **42,** 327–339.

Wilks, D. S., 1995: *Statistical Methods in the Atmospheric Sciences: An Introduction.* Academic Press, 467 pp.

Work, R. A., 1940: Standards and terminology. *Trans. Amer. Geophys. Union,* Part III, 979–981.

——, and R. T. Beaumont, 1958: Basic data characteristics in relation to runoff forecast accuracy. *Proc. Western Snow Conf.,* Bozeman, MT, Western Snow Conference, 45–53.